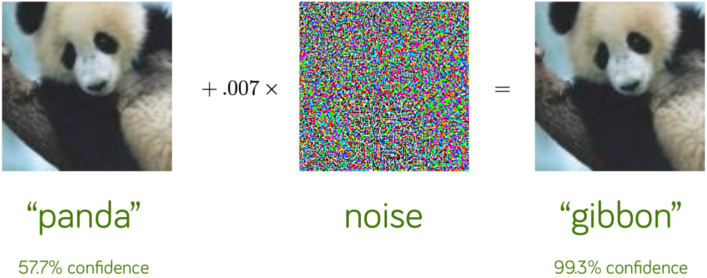
Formal Verification of Neural Networks

Lukas Koller and Tobias Ladner

Prof. Dr.-Ing. Matthias Althoff
Cyber-Physical Systems Group
Technische Universität München

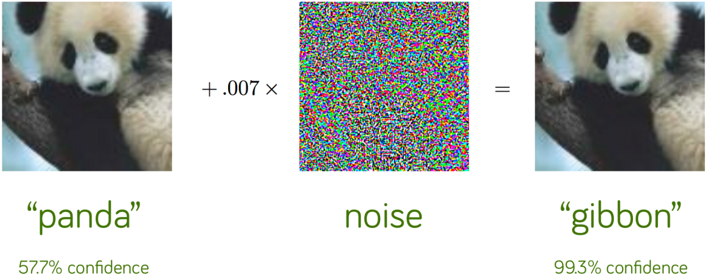
January 31st, 2024

Motivation



[1] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *International Conference on Learning Representations*. 2015.

Motivation

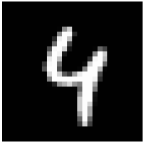


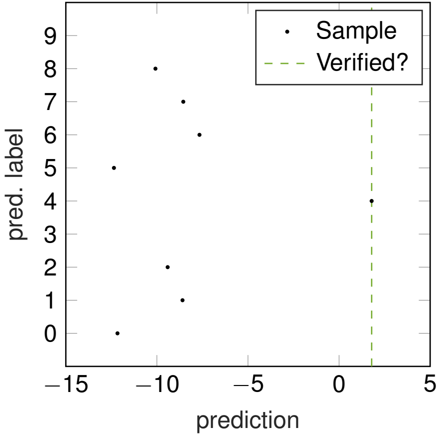
Adversarial examples[1] limit the applicability of neural networks in cyber-physical systems!

[1] Goodfellow, Shlens, and Szegedy, “Explaining and harnessing adversarial examples”.

Motivation

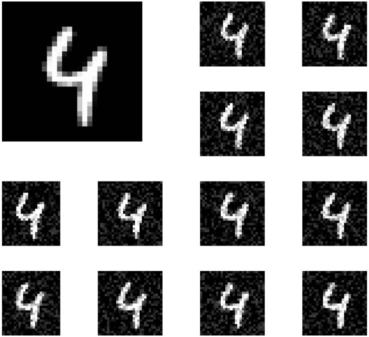
Let us demonstrate the formal verification of neural networks by an example:

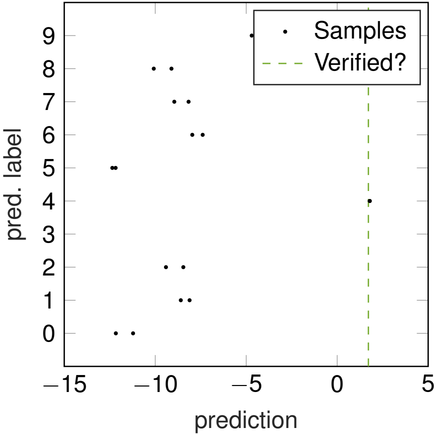




Motivation

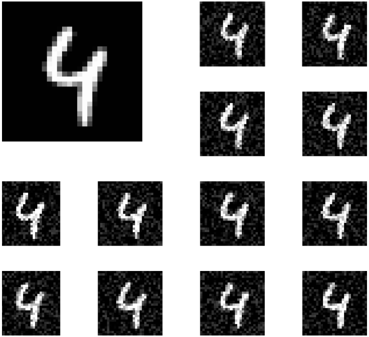
Let us demonstrate the formal verification of neural networks by an example:

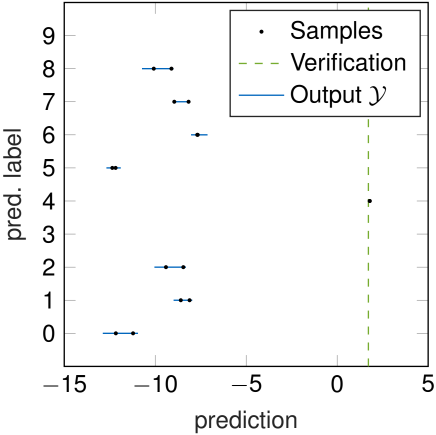




Motivation

Let us demonstrate the formal verification of neural networks by an example:





Your tasks:

- Learn about formal verification of neural networks
- Topics:
  1. Design network architectures for verification
  2. Implement and evaluate adversarial attacks
  3. Robust reinforcement learning

Your tasks:

- Learn about formal verification of neural networks
- Topics:
  1. Design network architectures for verification
  2. Implement and evaluate adversarial attacks
  3. Robust reinforcement learning

As an example, please check NNV in CORA[2]

Interested? Contact us!

Lukas Koller
lukas.koller@tum.de

Tobias Ladner
tobias.ladner@tum.de

[2] https://cora.in.tum.de/pages/neural-networks/